Why Might Negative Name Mover Heads Exist?

Callum McDougall, Arthur Conmy (*ARENA Interpretability Hackathon*)

Abstract

In this project we form a theory of why Negative Name Mover Heads (Wang et al., 2023) form in GPT-2 Small. We suspect that Negative Name Mover Heads i) respond to confident token predictions in the residual stream via Q-composition (Elhage et al., 2021), ii) attend to previous instances of such tokens in context and iii) negatively copy these tokens into the current token position. We use maximum activating dataset examples, negative copying score and a novel metric that tests our theory. Our results represent early research thoughts and are subject to ongoing investigation.[1]

For an up-to-date version of this writeup, we recommend reading our live report: https://www.overleaf.com/read/zfzrrppmmnyx

In this writeup we describe some background on Negative Name Mover Heads on the IOI distribution of text (Section 1), some evidence of the behavior of Negative Name Mover Heads (Section 2) and finally the most important contribution of our work thus far is a metric that tests our theory of Negative Name Mover behavior in any attention head in a model (Section 3). We find that indeed, the two Negative Name Mover Heads have the largest value of this metric in GPT-2 Small.

1 The IOI Distribution

We refer to Wang et al. (2023) for an introduction to and discussion of the Negative Name Mover Heads.

Why might we care about this pathological component in a language model? Prior work has found negative components to be a road-block to the automation of interpretability (Conmy et al., 2023). Additionally, Negative Name Mover Heads are related to the Backup Name Mover Heads found by Wang et al. (2023): the Appendix of that works shows that when mainline circuitry of a model is ablated, Negative Components can become positive components. Nanda (2023) discusses the problems with Backup Circuitry for the problem of *attribution* of things models do to internal components of those models. In future work we will document the strength of this evidence.

2 The General Distribution

We looked at which tokens the Negative Name Mover Heads pushed the most in the logits and which they pushed the least. We found the tokens they pushed the most were generally uninterpretable but the tokens they pushed the least were tokens that appeared in context and were often incorrect completions.

3 The Prediction-Attention Score

Based on the evidence from Section 2, we formed the hypothesis that Negative Name Mover Heads perform the following function:

[1] Our experiments are currently hosted at https://github.com/ArthurConmy/TransformerLens/tree/hackathon

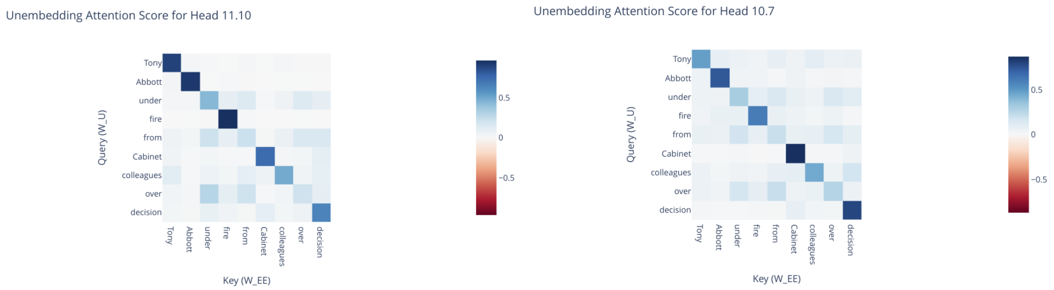

Figure 1: Prediction-Attention Score on the two Negative Name Movers

1. Respond to confident token predictions in the residual stream via Q-composition.
2. Attend to previous instances of such tokens (that the model is confident in predicting) in the current context.
3. Negatively copy these tokens into the current token position.

How might we test such a hypothesis? We observed that if 1. was carried out, we would expect that when query vectors are computed from the unembedding vectors for particular tokens, then the model attends to previous instances of that token in context. For example, let t be some token, $W_{E,t}$ be the embedding vector for that token, $W_{U,t}$ be the unembedding vector for that token and W_Q and W_K be the query and key parameters for a given head in a transformer language model. Then our hypothesis predicts that an attention-score-like quantity.

$$W_{E,s} W_K W_Q^T W_{U,t} \tag{1}$$

will be large when $s = t$ and smaller when $s \neq t$.

An issue with the setup described so far is that in GPT-2 Small ties the embeddings so $W_E = W_U$. Additionally, many prior works have found that Attention Layer 0 and especially MLP Layer 0 of GPT-2 Small act as 'effective embeddings' in that their outputs are mostly a function of the current token only (not mixing information across positions) and additionally the model appears to use their outputs to identify the token at a given position rather than the embedding matrix. Therefore we describe a way to generate an 'effective embedding matrix' W_{EE} in Appendix A.

Additionally, computing the attention paid between all 50 thousand tokens in GPT-2 Small's tokenizer is intractable, so we needed smaller subsets of tokens to search over to compare the $s = t$ and $s \neq t$ cases. We collect sets of distinct tokens B from OpenWebText documents and describe how we computed attention-score-like quantities for how much attention would be paid to various unembeddings. Specifically, we calculated the 'Prediction-Attention score', our novel metric for each of these bags of words B:

$$\text{Score}_{\text{Prediction-Attention}}(B) = \frac{1}{B} \sum_{t \in B} \text{Softmax}_{s \in B} \left[\left(W_{E,s} W_K W_Q^T W_{U,t} \right) \right]_t \tag{2}$$

Specifically, we take the softmax over the s dimension (varying the key vectors), and then we look at the tth entry of this distribution (t is used for the query). We find that the Prediction-Attention Score is very close to 1 for Negative Name Mover Heads.

In Figure 1 and Figure 2 we plot the 'attention' values for all queries and keys in a particular bag of tokens B. This corresponds to the green text in Equation (2). Therefore the Prediction-Attention Score is the average of the diagonal elements of these figures.

This compares favourably to almost all other heads in GPT-2 Small, such as S-Inhibition Heads and Positive Name Movers (Figure 2).

In fact, on random tokens the Prediction-Attention score was far larger for Negative Name Movers than almost all other heads (Figure 3).

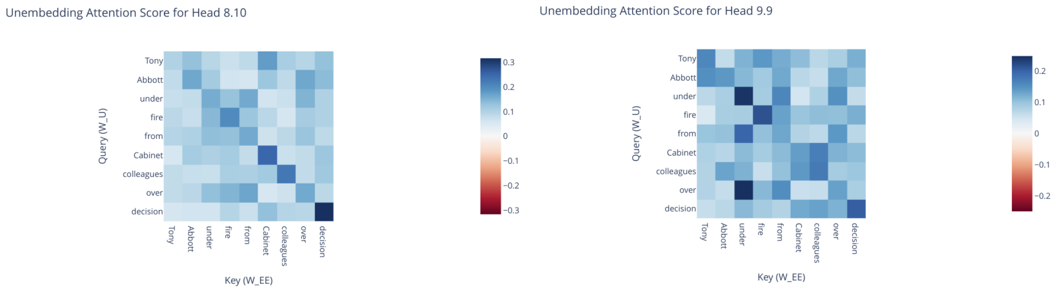

Figure 2: Prediction-Attention Score on Other Heads. See scale for low values

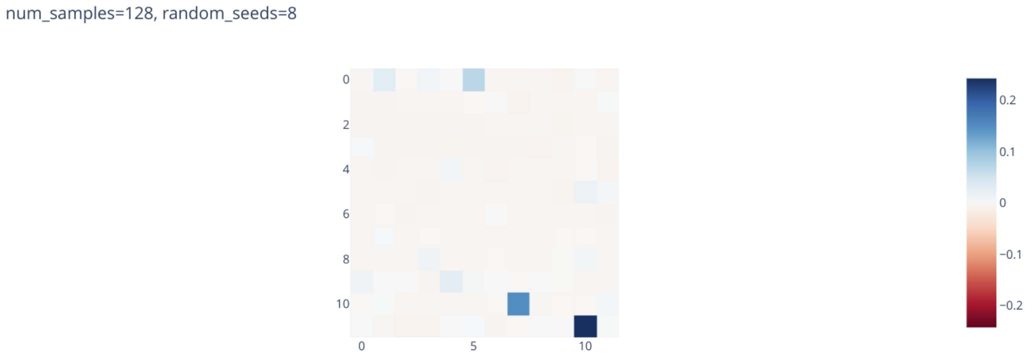

Figure 3: Prediction-Attention Score across GPT-2 Small's Heads

References

Wang, Kevin Ro, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt (2023). "Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small". In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=NpsVSN6o4ul.

Elhage, Nelson, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah (2021). "A Mathematical Framework for Transformer Circuits". In: *Transformer Circuits Thread*. URL: https://transformer-circuits.pub/2021/framework/index.html.

Conmy, Arthur, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso (2023). *Towards Automated Circuit Discovery for Mechanistic Interpretability*. arXiv: 2304.14997 [cs.LG].

Nanda, Neel (2023). *Open Problems in Mechanistic Interpretability*.

Appendix

A Effective Embedding

Suppose we set the attention pattern of all Layer 0 heads in GPT-2 Small to the identity matrix. Suppose we additionally set the positional embeddings of the model to all 0s. Then the output of the model at MLP 0 is solely a function of the input token. This corresponds to a different $d_{\text{vocab}} \times d_{\text{model}}$ matrix for computing the embeddings of a model that we call W_{EE}.